Recommend in a city with a higher population density and the best locality to opening a restaurant

Jhimy Cussi
May 22, 2021

1. Introduction

1.1. Background

Based on what has been learned, for this project it has been based on a fictitious client, the creation of the fictitious client is based on the inspiration of seeing many ventures that are born from a concern but do not have enough information to consolidate the business, especially in the line of restaurants. Therefore, the following lines will be based on the aforementioned.

From an idea of creating a business, specifically a restaurant as a consequence of the following criteria: The need to contribute to society by generating sources of employment and the availability of capital made up of the investment of several partners. However, investors have an urgent need to have sufficient certainty regarding the information on where to open the restaurant and that it can also be a profitable business.

1.2. Problem

Our client wishes to implement a restaurant in the state of New York, in order to determine the appropriate location that can guarantee the success of the business, our client considers the following aspects important: First, that the candidate cities must have a considerable population density; second, to guarantee the influx of the public (customers) to have options for appropriate restaurant categories and finally to determine if there is competition in the market, taking into account those options that do not generate much competition.

1.3. Interest

It is considered that the main stakeholders in this project are focused on the segment of investors or entrepreneurs who have in mind to open a restaurant, who seek information, it is considered that this project can contribute significantly in providing information on important variables that must be taken into account before making decisions and thus guarantee the success of the business.

1.4. Objective

Recommend the appropriate location to create a new restaurant in the New York state, in order to guarantee the success of the business.

2. Data acquisition and cleaning

2.1. Data sources

A data source has been sought which can provide us with sufficient information for this project, such as the most important variables such as latitude and longitude geospatial information, population density, state, city, etc. For this purpose, the following source of information has been considered:

Source: https://simplemaps.com/data/us-zips

Data: US Zip Codes Database

The fields that the dataset understands are listed below:

Field Name	Description
zip	The 5-digit zip code assigned by the U.S. Postal Service.
lat	The latitude of the zip code (learn more).
lng	The longitude of the zip code (learn more).
city	The official USPS city name.

Field Name	Description
state_id	The official USPS state abbreviation.
state_name	The state's name.
zcta	TRUE if the zip code is a Zip Code Tabulation area (learn more).
parent_zcta	The ZCTA that contains this zip code. Only exists if zcta is FALSE. Useful for making inferences about a zip codes that is a point from the ZCTA that contains it.
population	An estimate of the zip code's population. Only exists if zcta is TRUE.
density	The estimated population per square kilometer. Only exists if zcta is TRUE.
county_fips	The zip's primary county in the FIPS format.
county_name	The name of the county_fips.
county_weights	A JSON dictionary listing all county_fips and their weights (by population) associated with the zip code.
imprecise	TRUE if the lat/lng has been geolocated using the city (rare).
military	TRUE if the zip code is used by the US Military (lat/lng not available).
timezone	The city's time zone in the tz database format. (e.g. America/Los_Angeles)

Work with data such as zipcode, density, longitude, latitude, city and state, which will be complemented with the information provided by Foursquare that will allow us to determine or recommend the appropriate location to implement the restaurant in New York City.

2.2. Data cleaning

Data cleaning process, in this particular case "US Zip Codes Database", with the aim of eliminating noise in the information, generating unnecessary variation, for that purpose cases such as duplicate columns, duplicate rows and identification of incomplete data have been treated. or lost.

2.3. Feature selection

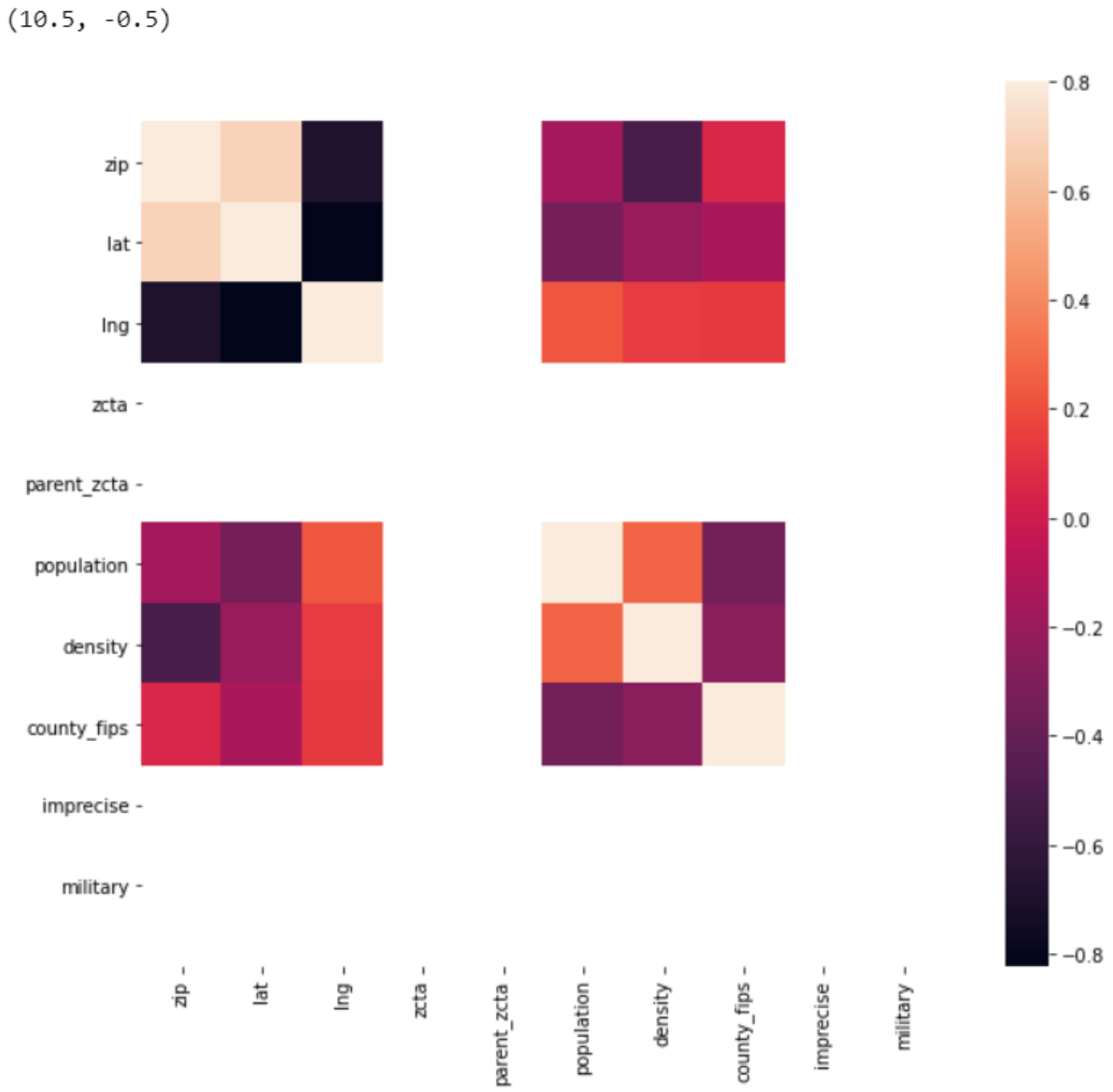
Descriptive statistics

Due to the restriction of this exercise, which is to take into account the population density, and also considering the descriptive statistics of the data, which helps us to consider the density and population fields.

	zip	lat	lng	parent_zcta	population	density	county_fips
count	200.00000	200.000000	200.000000	0.0	200.000000	200.000000	200.00000
mean	11007.97500	40.872002	-74.065843	NaN	43704.475000	16431.279000	36061.33000
std	900.15502	0.521784	0.851765	NaN	27168.485187	12260.765906	28.43084
min	10001.00000	40.551520	-79.049190	NaN	42.000000	3702.000000	36001.00000
25%	10309.50000	40.684083	-73.975142	NaN	22897.750000	5934.475000	36047.00000
50%	11204.50000	40.740815	-73.915095	NaN	38996.000000	12840.500000	36061.00000
75%	11373.25000	40.818865	-73.827967	NaN	63255.000000	23398.025000	36081.00000
max	14853.00000	43.150810	-73.125140	NaN	112088.000000	57641.100000	36119.00000

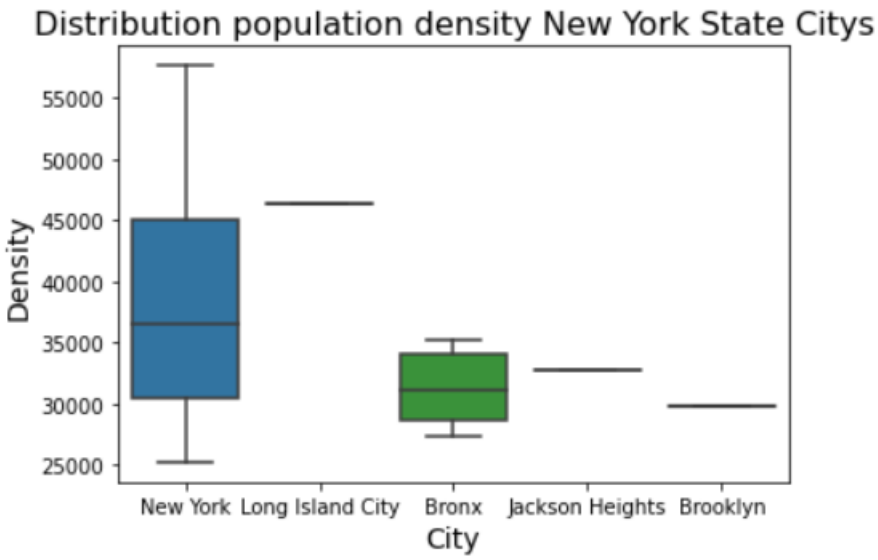
Correlation analysis

From the graph that determines the correlation between the variables on each axis, it can be interpreted, considering the density and population variables which will be subject to analysis, we can identify that their correlation is low.



Distribution analysis

From the graph we can interpret the dispersion with respect to the IQR (Inter-Quartile Range) that measures the variability of 50% of the data, specifically considering New York and Bronx (which can be seen), it can be interpreted that the variability of New York it is greater than the variability of the Bronx. Regarding the median which represents the distribution of the data, it can be said that New York has a positive asymmetric distribution or positively skewed upwards, with respect to Bronx, it can be said that the median is almost in the center therefore its data distribution is nearly symmetric.



3. Methodology

In the article titled "Understanding and projecting the restaurantscape: Influence of neighborhood

sociodemographic characteristics on restaurant location ", August 2017, written by Yang Yang and Jing-Huei Huang, at: https://www.researchgate.net/publication/318813299, mention that to understand the location patterns of different types of restaurants In the United States, they investigated the relationship between sociodemographic characteristics, in which they reflect the impact generated by sociodemographic factors such as population density, middle age of the population, educational level, etc., taking into account the aforementioned as a theoretical basis, In this exercise, the density of the population and population will be taken into account, considering the variables that our data comprises.

After having the data "US Zip Codes Database", and after having taken into account the state "New York" up to this point we have already processed:

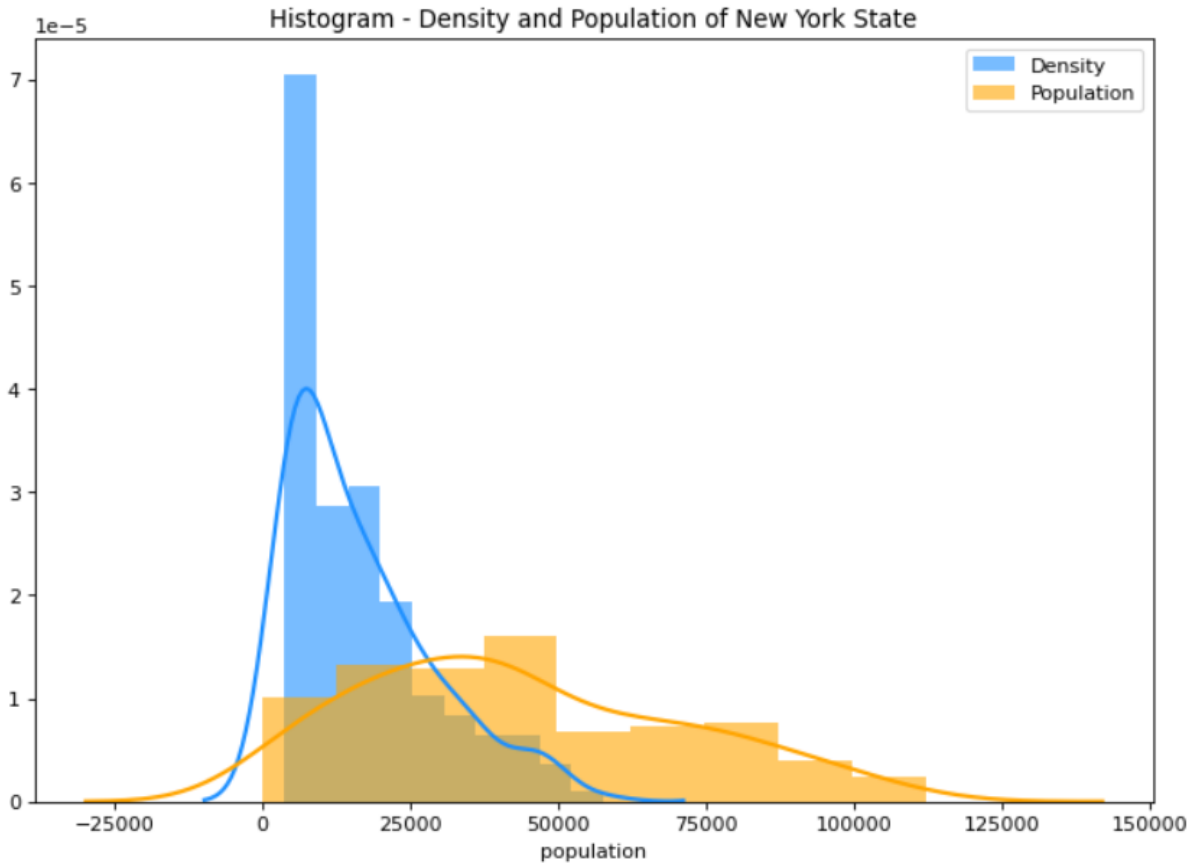
- The first 200 cities with the highest population density in order to determine the recommended location to open the restaurant.

- Now we will proceed with other activities such as:
- Obtain information from the Forsquare API to obtain up to 100 existing restaurants within a radius of 500 meters.
- To contribute to the analysis, classify restaurants by category to determine frequency by city.
- Determine the 10 cities with the highest population density, in order to subsequently select the cities that are within the top 10.
- We will focus on the most promising areas and within them we will create groups of locations focused on recommending the best place to implement a restaurant: After obtaining up to 100 restaurants within a radius of 500 meters, we will later present a map of all those locations.
- Determine the top 10 restaurant categories for each city and town.
- We will proceed with the cluster applying k-means grouping of those locations to identify the locality, for the exploitation of the optimal location where we will recommend the implementation of a restaurant.
- To recommend, the k-means result should be restricted to the cities with the highest population density.

4. Analysis

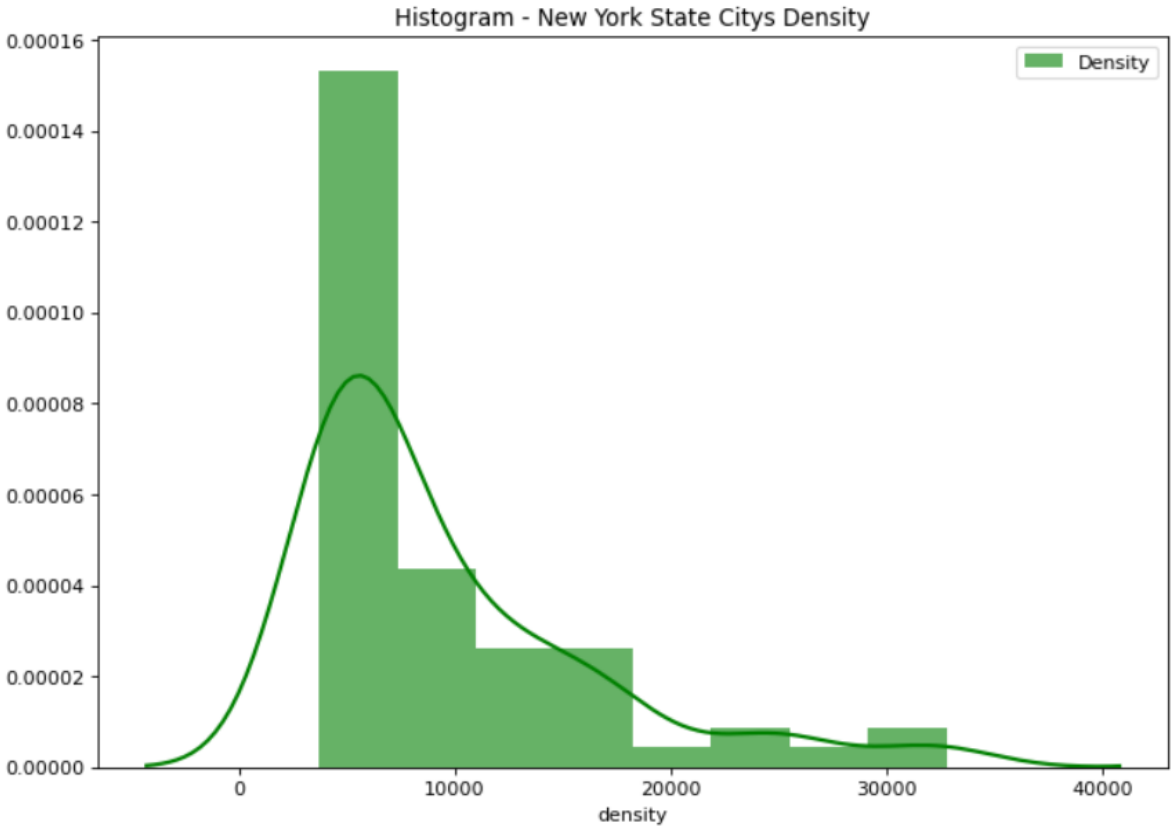
4.1. New York State density and population distribution

The visualization of the distribution of the density and the population of the state of New York is carried out, in order to validate the requirement of the clients, who demand that a restaurant be created in cities with higher population density and therefore a considerable population.



4.2. New York State Citys density

After validating the population density and population of the state of New York, only the population density can be considered by the shape of the distribution, which reveals the existence of cities that have a significant greater population density. Next, the distribution of the population density of the cities of the state of New York is visualized and analyzed.

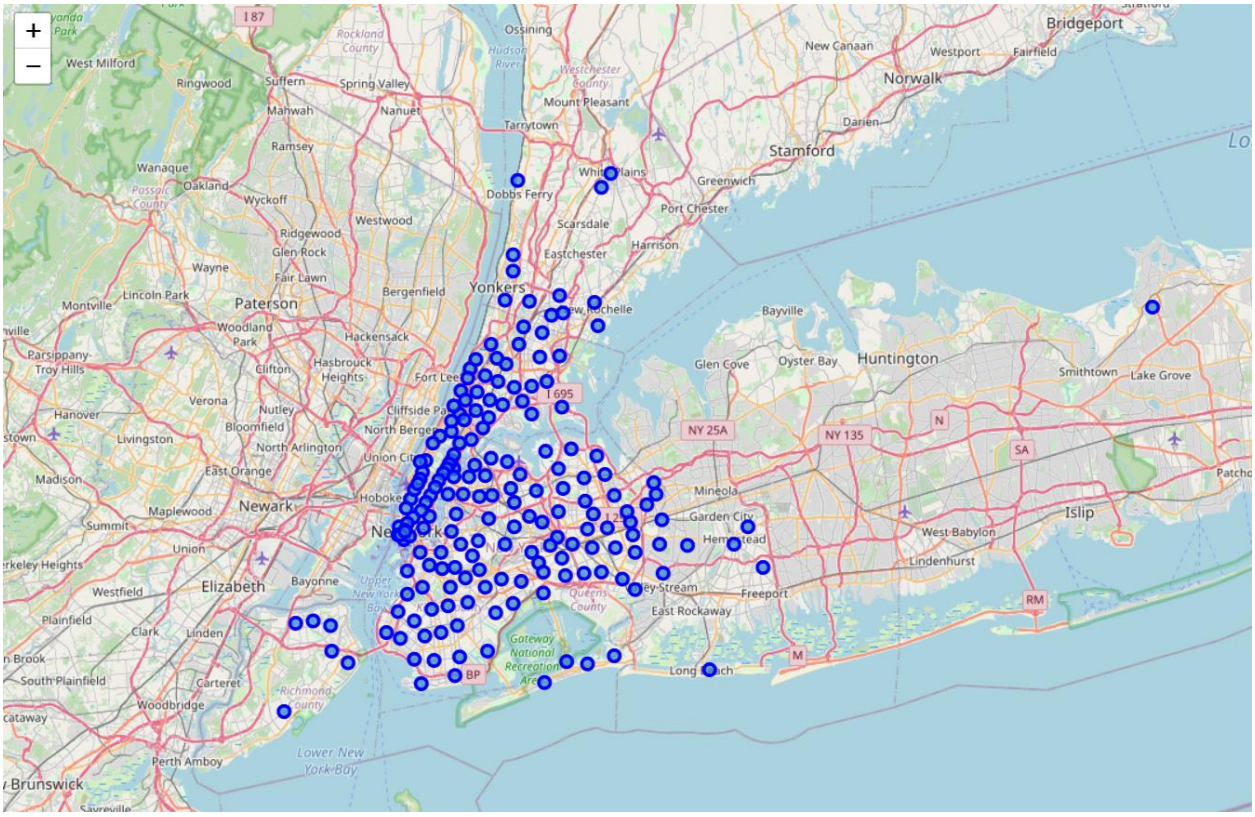


4.3. Get data from API Forsquare - The 100 restaurants within a radius of 500 kilometers

Get data from API Forsquare with the information from database "US Zip Codes Database".

	location	latitude	longitude	venue	venue_latitude	venue_longitude	venue_category
0	New York	40.7763	-73.95372	Levain Bakery	40.777354	-73.955284	Bakery
1	New York	40.7763	-73.95372	Elio's	40.776783	-73.952736	Italian Restaurant
2	New York	40.7763	-73.95372	Luke's Lobster	40.774805	-73.954423	Seafood Restaurant
3	New York	40.7763	-73.95372	Schaller's Stube Sausage Bar	40.777588	-73.951975	Hot Dog Joint
4	New York	40.7763	-73.95372	Heidelberg Restaurant	40.777532	-73.951979	German Restaurant
5	New York	40.7763	-73.95372	San Matteo Pizzeria e Cucina	40.774674	-73.954221	Italian Restaurant
6	New York	40.7763	-73.95372	The Penrose	40.775444	-73.953143	Gastropub
7	New York	40.7763	-73.95372	Italianissimo Ristorante	40.776351	-73.952179	Italian Restaurant
8	New York	40.7763	-73.95372	H&H Midtown Bagels East	40.774446	-73.954479	Bagel Shop
9	New York	40.7763	-73.95372	Flex Mussels	40.776337	-73.956430	Seafood Restaurant

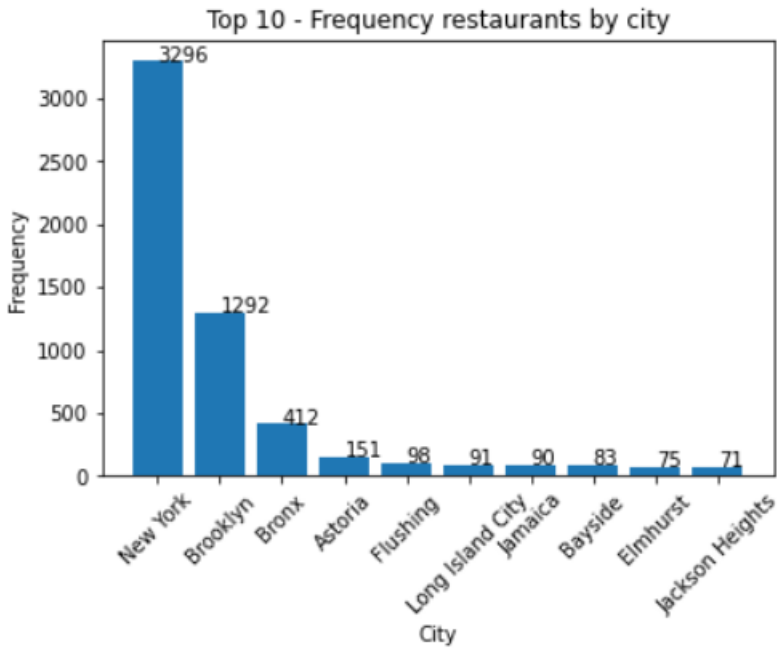
Below you can see the geospatial distribution of the information:



4.4. Frequency of restaurants by city based on "Foursquare" data

The top 10 is determined, the top 10 with high frequency of restaurants by city.

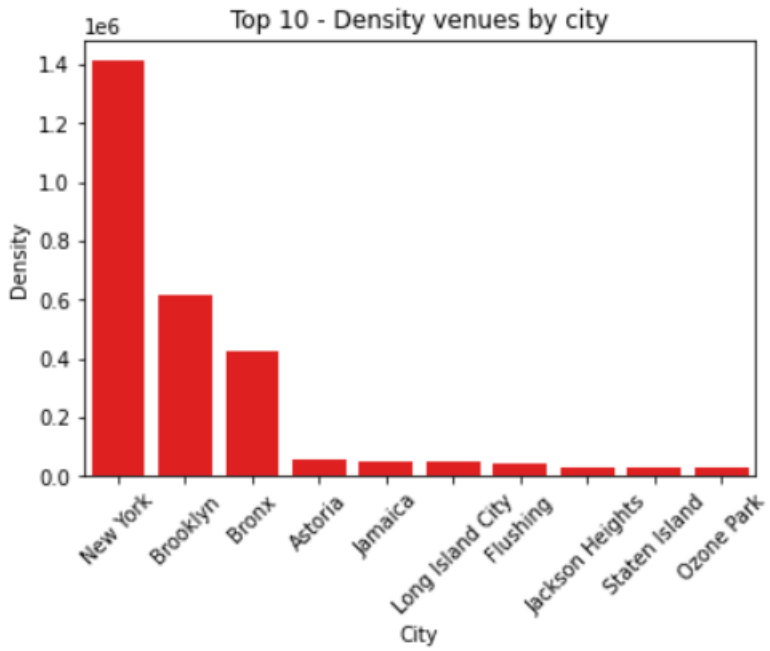
	location	freq
35	New York	3296
7	Brooklyn	1292
6	Bronx	412
2	Astoria	151
17	Flushing	98
30	Long Island City	91
27	Jamaica	90
4	Bayside	83
13	Elmhurst	75
26	Jackson Heights	71



4.5. The highest population density from data "US Zip Codes Database"

Obtaining the first 10 cities with the highest population density, they consider that it is intended to open the restaurant under this restriction.

	location	density
36	New York	1412008.4
8	Brooklyn	617245.1
7	Bronx	426312.4
3	Astoria	59608.6
28	Jamaica	52757.7
31	Long Island City	51045.9
18	Flushing	41525.7
27	Jackson Heights	32758.7
53	Staten Island	29333.0
39	Ozone Park	26411.0



This information will be chosen to process in the population density restriction, because the origin of the information source is official.

5. Modeling

K-means is an unsupervised classification algorithm that groups objects into k groups based on their characteristics, the groupings are performed by minimizing the sum of distances between each object and the centroid of its subset or cluster.

There are many types of clustering algorithms, such as partitioning, hierarchical grouping or density-based, in this particular case K-Means is a type of partition grouping, which means that it divides the data into k subsets or groupings that are not they overlap and also without any structure or internal grouping labels. The objects within a cluster (subset) are very similar, and the objects in different clusters are very different. Therefore, it means that it is an unsupervised algorithm.

6.1. Clustering

In this section, the type of restaurant and the 10 most representative will be taken into account, therefore we proceed by grouping rows by location and taking the mean of the frequency of occurrence of each category, preparing the data frame for grouping.

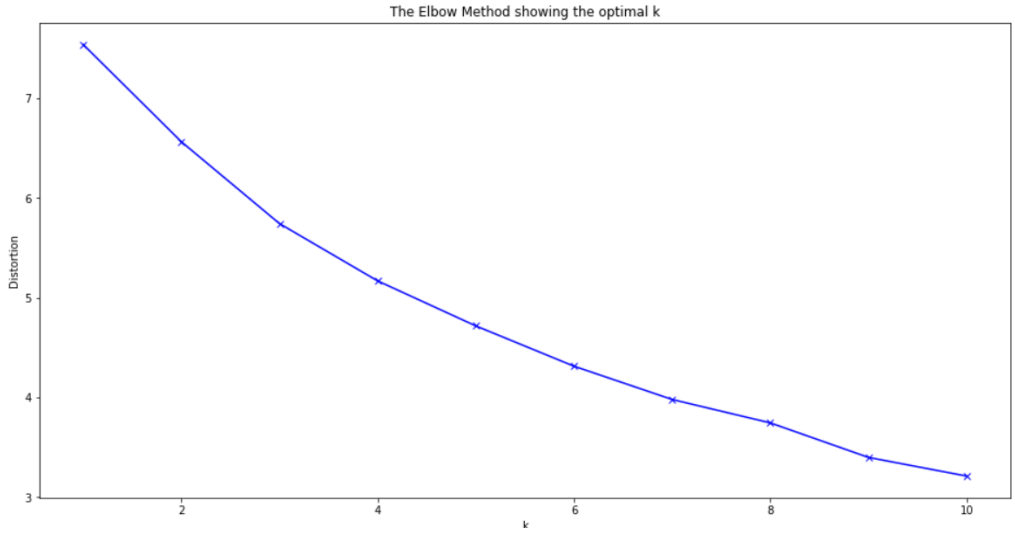
It is important to consider the transformation of the information collected using the one-hot encoding method.

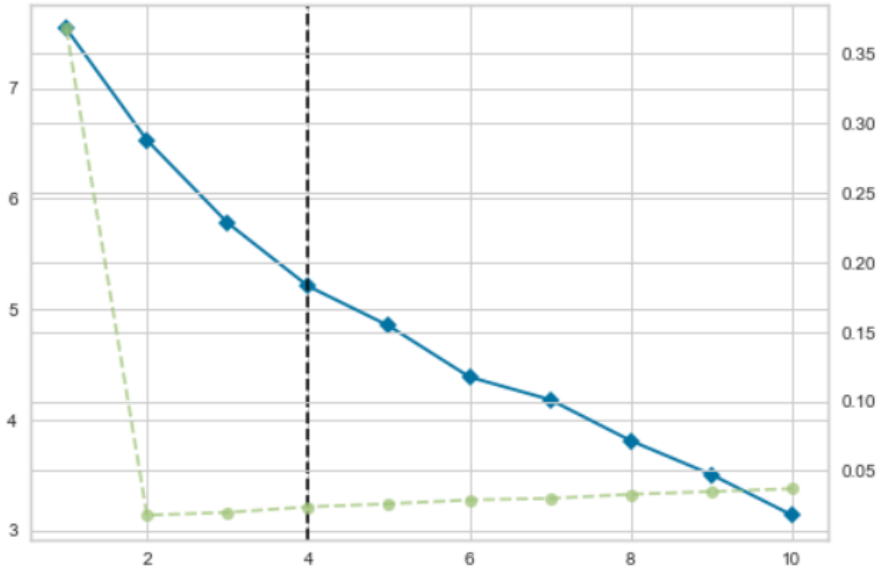
	location	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	...
0	Albany	0.000000	0.0	0.000000	0.000000	0.000000	0.076923	0.0	0.0	0.000000	...
1	Arverne	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	...
2	Astoria	0.013245	0.0	0.000000	0.006623	0.000000	0.000000	0.0	0.0	0.006623	...
3	Auburn	0.000000	0.0	0.125000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	...
4	Bayside	0.000000	0.0	0.048193	0.000000	0.000000	0.024096	0.0	0.0	0.000000	...
5	Bellerose	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.000000	...
6	Bronx	0.000000	0.0	0.016990	0.000000	0.000000	0.009709	0.0	0.0	0.004854	...
7	Brooklyn	0.000000	0.0	0.020898	0.001548	0.000774	0.016254	0.0	0.0	0.001548	...
8	Buffalo	0.000000	0.0	0.030303	0.000000	0.000000	0.030303	0.0	0.0	0.000000	...
9	Cambria Heights	0.000000	0.0	0.000000	0.000000	0.000000	0.111111	0.0	0.0	0.000000	...

Getting the best "k" applying elbow method

The "elbow" method is a common heuristic in mathematical optimization, allowing you to select a point at which diminishing returns no longer justify the additional cost, being an inflection point in the ignorable decreasing part. This selection makes it possible to provide a much better model for the data.

With the "elbow" method, the optimal number of clusters is selected by adjusting the model with a range of values for K, which consists of checking the inflection point in the curve, this parameter definitely applicable to the model for its best fit.



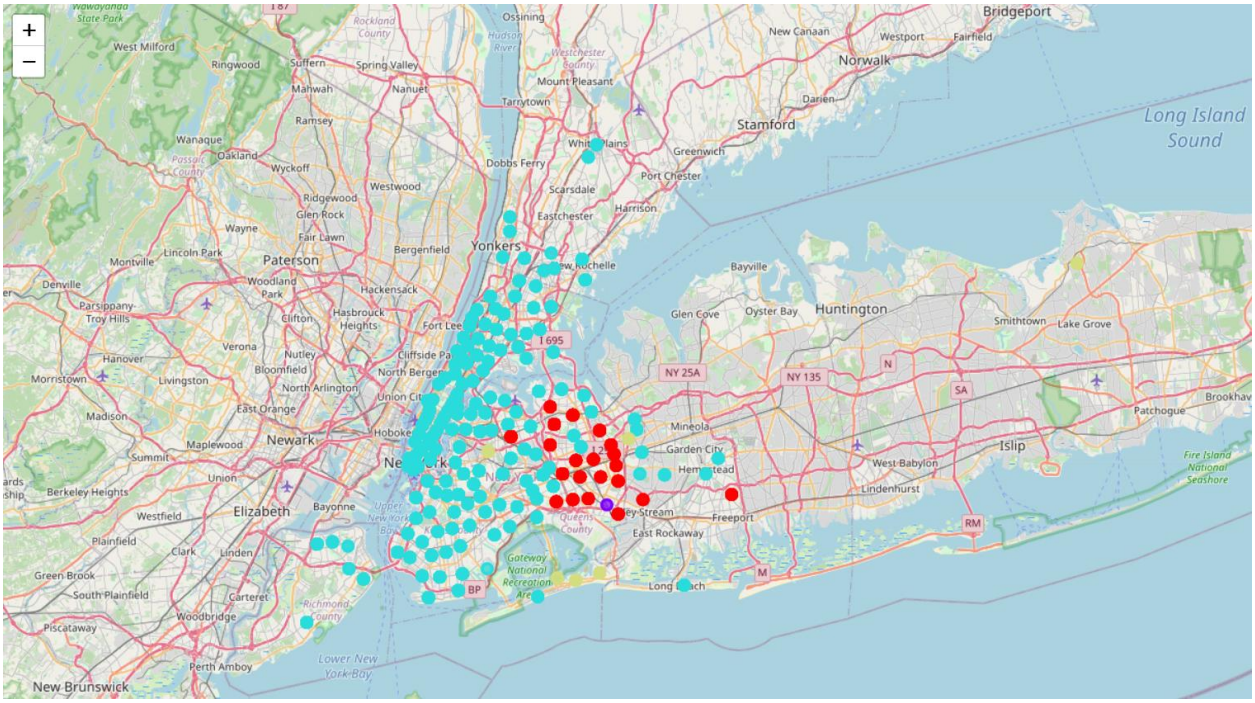


The value of k at the “elbow”, the point after which the distortion/inertia start decreasing in a linear fashion. Therefor the value k=4

Running processing with K-means

After finding the recommended k, the execution (fit) proceeds applying K-means with k=4, resulting after segmentation into 4 subsets or clusters.

Finally, let's visualize the resulting clusters:



Labels:

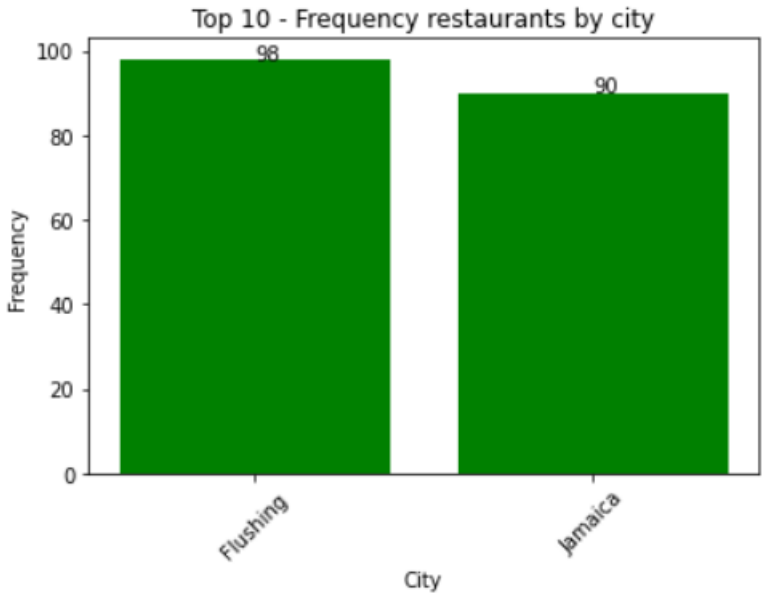
- **Red:** Cluster 0
- **Purple:** Cluster 1
- **Light blue:** Cluster 2
- **Green:** Cluster 3

6.2. Result

To the result of K-means, the population density restriction will be added, considering a study variable in the present exercise, in order to have information that allows us to recommend from a limited result.

For all the K-means results, it has been restricted to the cities with the highest density, thus limiting as much as possible the probabilities of locations where it is convenient to recommend for the opening of the restaurant, in addition to recommending the appropriate categories. The results by cluster are described below.

Cluster 1:



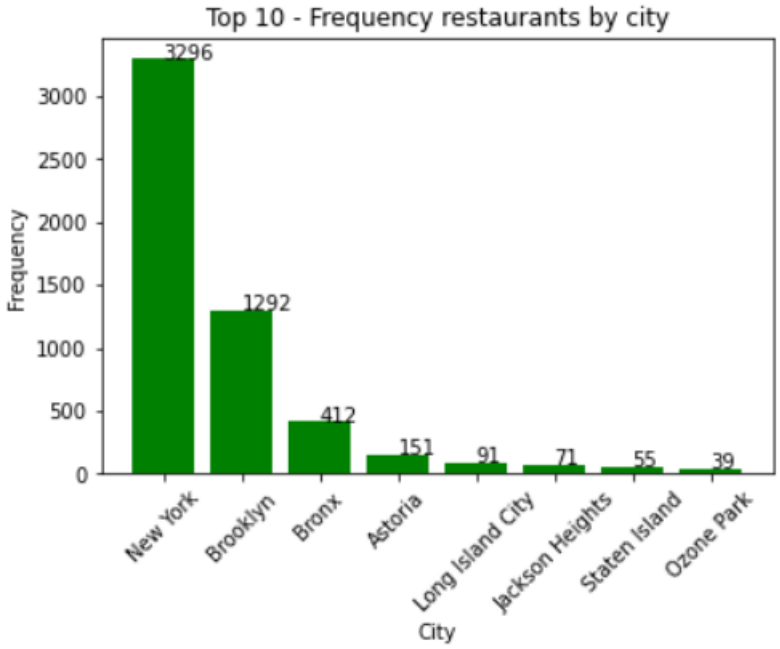
	location	latitude	longitude	venue	venue_latitude	venue_longitude	venue_category	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
count	188	188.000000	188.000000	188	188.000000	188.000000	188	188	188	188	188	188	188	188	188	188	188
unique	2	NaN	NaN	172	NaN	NaN	48	1	2	1	2	2	2	2	2	2	2
top	Flushing	NaN	NaN	Dunkin'	NaN	NaN	Chinese Restaurant	Chinese Restaurant	Korean Restaurant	Pizza Place	Bakery	Indian Restaurant	Deli / Bodega	Food Truck	Asian Restaurant	Donut Shop	Fried Chicken Joint
freq	98	NaN	NaN	8	NaN	NaN	32	188	98	188	98	98	98	98	98	98	98
mean	NaN	40.726793	-73.808430	NaN	40.726744	-73.808328	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	0.029916	0.014079	NaN	0.030160	0.014221	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	40.675770	-73.827390	NaN	40.672979	-73.832444	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	40.701380	-73.820920	NaN	40.701650	-73.821755	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	40.730410	-73.809680	NaN	40.730131	-73.808877	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	40.751480	-73.796320	NaN	40.753468	-73.795812	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	40.768630	-73.776260	NaN	40.771920	-73.775162	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

In the result of the cluster, it has been restricted to the 10 cities with the highest population density, which is reflected to the cities "Flushing" and "Jamaica" with a frequency of 96 and 90 respectively. Later you can see the statistical result where the "top" (most common data) turn out to be the city "Flushing" and the locality "Dunkin '". From the previous result we also have the 10 most common restaurant categories, of which the last 3 are taken into account to mitigate the competition, but at the same time ensuring that it is within the top 10; as a result, there are 8th- "Asian Restaurant", 9th- "Donut Shop" and 10th- "Fried Chicken Joint".

Cluster 2:

In the result of the cluster, it has been restricted to the 10 cities with the highest population density, of which there are no cities that are within the 10 cities with the highest density, for which there are no results to recommend.

Cluster 3:



	location	latitude	longitude	venue	venue_latitude	venue_longitude	venue_category	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
count	5407	5407.000000	5407.000000	5407	5407.000000	5407.000000	5407	5407	5407	5407	5407	5407	5407	5407	5407	5407	5407
unique	8	NaN	NaN	4155	NaN	NaN	133	5	3	4	6	7	7	8	8	8	7
top	New York	NaN	NaN	Dunkin'	NaN	NaN	Pizza Place	Italian Restaurant	Pizza Place	Deli / Bodega	Mexican Restaurant	Café	American Restaurant	Restaurant	Bakery	Chinese Restaurant	Sandwich Place
freq	3296	NaN	NaN	104	NaN	NaN	431	3296	4924	3387	3387	3296	3387	3296	3296	3296	4588
mean	NaN	40.739140	-73.965449	NaN	40.739087	-73.965421	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	NaN	0.066198	0.043081	NaN	0.066055	0.042960	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	NaN	40.551520	-74.150380	NaN	40.549539	-74.151117	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	NaN	40.706140	-73.996360	NaN	40.707288	-73.994147	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	40.741870	-73.967930	NaN	40.742826	-73.970413	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	40.776010	-73.943960	NaN	40.774919	-73.944230	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	NaN	40.895940	-73.822980	NaN	40.900073	-73.818528	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

In the result of the cluster, it has been restricted to the 10 cities with the highest population density, which is reflected to the cities "New York", "Brooklyn", "Bronx", "Astoria", "Long Island City", "Jackson Heights "," Staten Island "and" Ozone Park "with frequencies of 3296, 1292, 412, 151, 91, 71, 55, 39 respectively. Later you can see the statistical result where the "top" (most common data) turn out to be the city "New York" and the locality "Dunkin '". From the previous result we also have the 10 most common restaurant categories, of which the last 3 are taken into account to mitigate the competition, but at the same time ensuring that it is within the top 10; as a result, we have 8th- "Bakery", 9th- "Chinese Restaurant" and 10th- "Sandwich Place".

Cluster 4:

In the result of the cluster, it has been restricted to the 10 cities with the highest population density, of which there are no cities that are within the 10 cities with the highest density, for which there are no results to recommend.

As a result, the recommendation table is presented below:

Recommendation table

Cluster	City	Venue	Restaurant Category
Cluster 1	Flushing	Dunkin'	*Asian Restaurant -> Donut Shop -> Fried Chicken Joint*
Cluster 2	NaN	NaN	*NaN*
Cluster 3	New York	Dunkin'	*Bakery -> Chinese Restaurant -> Sandwich Place*
Cluster 4	NaN	NaN	*NaN*

6. Conclusion

The objective of the project is to recommend the appropriate location to create a new restaurant in New York State, in order to guarantee the success of the business. Besides, it is also recommended that types of restaurants are suitable options, considering the restriction on the 10 cities with the highest population density. After finishing the corresponding stages of the project, it is concluded that our client can open his restaurant, taking into account the following recommendation:

State: New York

Cities: Flushing, New York

Types of restaurant:

In the city of "Flushing":

- 1st. option: "Fried Chicken Joint"
- 2nd. option: "Donut Shop"
- 3rd. option: "Asian Restaurant"

In the city of "New York":

- 1st. option: "Sandwich Place"
- 2nd. option: "Chinese Restaurant"
- 3rd. option: "Bakery"

7. Future directions

In the present work, only some other variables that are also important have been considered. It can be investigated and analyzed with other sociodemographic variables such as the level of education, income and others that allow to determine a greater precision when selecting the best options of locations where to open a restaurant; From something as logical as what specialty those interested in opening a restaurant have, this aspect would define the type of restaurant, other variables can also be considered such as the existence of markets for the supply of inputs, the proximity of institutions or companies, universities, etc. which reveals the fluidity and concentration of people at certain times. The variables are only a few that have been mentioned, surely this work can be the beginning of an interesting investigation.

8. References

- https://www.scikit-yb.org/en/latest/api/cluster/elbow.html
- https://www.linkedin.com/pulse/applied-data-science-capstone-project-restaurant-wagner-mba
- https://towardsdatascience.com/strategic-location-for-establishing-an-asian-restaurant-c3aecf2496b1
- https://magnet.xataka.com/en-diez-minutos/verdad-densidad-poblacion-no-importa-total-sino-densidad-habitada-1